# What can the genetic data tell us about the history?

**Vladimir Voevodsky**

Started April 14, 2004

**Introduction** There are two kinds of histories. Histories of the first kind concentrate on a small number of key individuals and interpret the past events through their relation to these individuals. Histories of the second kind are concerned more with the dynamics of large groups of people - migrations, expansions or contractions of whole populations, changes in customs, culture and technology etc. The modern advances in our ability to extract information from the genomes can help to construct booth kinds of histories. Here we are concerned mostly with the histories of the second kind.

**The models** In order to use quantitative methods to reason about history we need to have a class of models. Our main objects of interest are people or groups of people and events which occur to them. Following the demographic approach we fix a set $C$ of *cohorts*, or types of members of the population, which we distinguish in our model. For a given $C$, a $C$-structured population is a finite set $P$ together with a map $P \to C$. This map defines a decomposition

$$P = \coprod_{c \in C} P_c$$

where $P_c$ is the fiber of $P \to C$ over $c$. For a $C$-structured population $P$ the formal linear combination

$$\underline{c}(P) = \sum_{c \in C} n_c(P) c$$

where $n_c$ is the number of elements in $P_c$, is called the cohort composition of $P$. Cohort compositions $\sum n_c c$ can be identified with elements of the set

$$S^\bullet C := \coprod_{n \geq 0} S^n C$$

where $S^n C$ is the n-th symmetric power of $C$.

In the most simple case, $C$ is a point and the cohort composition is determined by the "multiplicity" of this point which equals to the size of the population. In a more complicated case there may be two types say juveniles and adults $C = \{j, a\}$ and the composition is determined by two numbers $n_j(P)$ and $n_a(P)$ - the number of juveniles and adults respectively. In more realistic models $C$ may have many elements reflecting age, location, occupation, genetic or phenotypic markers etc. Members of certain types may not correspond to people (at least to individual people) at all. In particular it seems to be important to be able to consider models where there is a cohort whose members are married couples or possibly families. One may also consider the case when $C$ contains $\mathbf{R}_+$ corresponding to the precise age or $S^2$ corresponding to the precise location on earth. It is not clear to me at the moment whether such continuous models have any advantage over their discrete approximations. In what follows I assume $C$ to be finite or at most countable.

We further fix a set $T$ of types of *events* and a map $T \to S^{\bullet}C \times S^{\bullet}C$ which assigns to an event type $\tau$ its signature $(el(\tau), pr(\tau))$. An event with signature $(el(\tau), pr(\tau))$ transforms a sub-population with cohort composition $el(\tau)$ to a sub-population with cohort composition $pr(\tau)$. For example, the death of a member of cohort $c$ is an event with signature $(c, 0)$ while, say, a formation of a couple by members of cohorts $f$ and $m$ is an event of type $(f + m, c)$ where $c$ is the cohort of couples. We say that a sub-population is eligible for an event of type $\tau$ if its cohort composition coincides with the one specified by the first half of the signature of $\tau$ and that it is produced by an event of type $\tau$ if its cohort composition is the one specified by the second half of the signature of $\tau$. Migrations, births, divorces, changes in occupation, mutations etc. are all possible types of events in our models.

The pair of sets $C$ and $T$ together with the signature map $sig : T \to S^{\bullet}C \times S^{\bullet}C$ is called a historic scheme.

Let $S$ be a historic scheme and $[t_0, t_1]$ a time interval. An $S$-history during $[t_0, t_1]$ is the following collection of data:

1. a finite (possibly empty) sequence $s_1, \ldots, s_n \in [t_0, t_1]$ such that $t_0 < s_1 < \ldots < s_n < t_1$
2. for any $i = 0, \ldots, n$ a finite set $P_i$ together with a map $P_i \to C$
3. for any $j = 1, \ldots, n$ a finite set $E_j$ together with a map $\tau : E_j \to T$
4. for any $j$ and any $e \in E_j$ two subsets $s(e) \subset P_{j-1}$ and $t(e) \subset P_j$ such that
   (a) for $e' \neq e$ one has $s(e) \cap s(e') = \emptyset$ and $t(e) \cap t(e') = \emptyset$
   (b) $\underline{c}(s(e)) = el(\tau(c))$ and $\underline{c}(d(e)) = pr(\tau(e))$
5. for any $j$ a bijection
$$[\mathbf{eq1}] P_{j-1} - \cup_{e \in E_j} s(e) \to P_j - \cup_{e \in E_j} t(e) \tag{1}$$

An $S$-history can be described as follows. We have a set of moments of time $s_1, \ldots, s_n$ when the events happen. The complement to this set is a union of $n + 1$ "quiet" intervals. During each of these intervals we have a well defined $C$-structured population $P_j$. For each $i$, $E_i$ is the set of events which occurred at $s_i$ with the map $E_i \to T$ providing the types of these events. For an event $e$, $s(e)$ is the smallest sub-population which was affected by this event and $t(e)$ the new sub-population produced from it. The bijection (1) provides an identification between the members of the populations $P_{j-1}$ and $P_j$ which were not affected by the events in $E_j$.

We say that two $S$-histories $H$ and $H'$ on the same time interval are equivalent (or isomorphic) if the corresponding sets of event times coincide and there are bijections between the sets $P_i$ (resp. $E_j$) and $P'_i$ (resp. $E'_j$) preserving all the structures. Let $h(t_0, t_1; S)$ be the set of equivalence classes of $S$-histories on $[t_0, t_1]$.

For a given $S = (C, T, sig)$ and a given interval of time $[t_0, t_1]$ we define historic conditions as a function $\rho : T \times (t_0, t_1) \to \mathbf{R}_{\geq 0}$. Intuitively, the value of this function on $(\tau, t)$ is the rate of occurrence of events of type $\tau$ at $t$ normalized with respect to the number of sub-populations

eligible for $\tau$. For example if we have cohorts $f$, $m$ and $c$ understood as males, females and couples and we have an event type $\tau$ of signature $(m + f, c)$ with the meaning of marriage then

$$\rho(\tau, t) = lim_{\Delta t \to 0} P(\tau, [t, t + \Delta t])/\Delta t$$

where $P(\tau, [t, t + \Delta t])$ is the probability that a given male and a given female will marry during the interval $[t, t + \Delta t]$.